面向搜索引擎查询日志的领域术语自动识别方法*

刘 彤 倪维健 柳 梅
(山东科技大学信息科学与工程学院 青岛 266590)

摘要: 【目的】为弥补传统基于静态领域语料的领域术语识别方法的不足, 提出一种从搜索引擎查询日志中自动识别领域术语的新方法。【方法】使用四部图对查询日志进行抽象描述, 并在其上应用流形排序算法得到所有候选术语关于领域度的排序, 取排在前列的术语作为领域术语。【结果】在真实搜索引擎的查询日志上实验证实本文方法具有更好的领域术语识别效果, 在 Precision@n 指标上比基准方法提升约 20%。【局限】识别到的领域术语的覆盖面部分依赖于领域专家选取的初始查询词, 这对领域专家的经验提出一定要求。【结论】该方法无需事先准备大规模领域语料以及大量的人工标注, 即可构建高质量的领域术语集合, 具有较高的实用价值。

关键词: 领域术语 搜索引擎 查询日志 流形排序
分类号: TP391.1

1 引 言

领域术语泛指经常出现于特定领域语料中的短语[1], 比如"双条杉天牛"和"叶斑病"是农业领域的术语。相比于传统的通用词汇, 领域术语蕴含丰富的领域知识, 因而领域词典成为各类情报处理与分析任务的一类基础资源。现有领域字典主要通过两种方式构建: 采用手工构建, 如 AGRIVOC[2]、UMLS[3]等, 这种方法虽然准确率较高, 但构建过程耗费大量人力, 特别是在领域知识更新时很难对领域词典进行有效维护; 从领域新闻[4]、科技文献[5-7]、维基百科[8]、专利文档[9-10]、领域网站[11]等领域语料中自动识别领域术语, 这种方法虽然具有一定的自动化程度, 但是识别效果很大程度取决于领域语料的质量。一般而言, 获取高质量领域语料存在一些现实困难: 一方面, 同时具有较大领域覆盖面和较强领域相关性的领域语料通常难以采集; 另一方面, 领域语料通常是静态的, 更新频度较低, 难以适应领域知识的最新发展变化。这使得基于领域语料的领域术语识别方法在现实应用中面临着很大挑战。

与传统的领域语料相比, 搜索引擎查询日志是一种新型的语料资源[12], 它由搜索引擎自动采集, 记录用户与搜索引擎的整个交互过程, 包括用户提交的查询词、搜索时间、搜索结果、用户点击等。查询日志具有如下特点:

(1) 海量性: 搜索引擎的广泛应用使其积累了海量的查询日志, 不仅数量庞大, 而且覆盖面广, 基本涵盖了各领域的信息需求;

(2) 实时性: 查询日志是实时更新的, 记录了每条用户最新提交的查询请求, 因而能够反映各领域最新的信息需求。

上述特点使得查询日志蕴含了丰富的领域术语, 从而可以被当作识别领域术语的一类重要数据资源。由于查询日志由搜索引擎自动采集, 因而无需事先准

通讯作者: 倪维健, ORCID: 0000-0002-7924-7350, E-mail: niweijian@gmail.com。

*本文系山东省自然科学基金"动态环境下结构支持向量机学习算法及其应用研究"(项目编号:ZR2014FP011)、山东省高等学校科技计划项目"面向信息检索的非平衡数据排序学习问题研究"(项目编号:J12LN45)和山东省高等学校科技计划项目"面向非规范分布形态下不平衡文本数据的监督学习关键技术研究"(项目编号:J14LN33)的研究成果之一。

备高质量领域语料即可从查询日志这种非领域语料中自动识别高质量的领域术语集合，在现实应用中具有更高的推广价值。

2 国内外研究现状

2.1 领域术语识别

领域术语识别是情报学科的一个经典问题[13]，在整体上可以分为三类：基于规则、基于统计和基于机器学习的方法。

基于规则的方法通过领域专家的参与，利用构词学规则、语义或词性信息构造模板，通过模板匹配识别领域术语[5,14]，其主要不足在于编写和维护规则费时费力，而且随着语料集的增大，规则的完备性越来越难以保证。为克服该问题，提出基于统计的领域术语识别方法，这类方法基于候选词条在语料中的统计特征进行领域术语识别，代表性的有TF-IDF指标[6]、互信息指标[9]等。研究者还提出多种复合指标，比如Dorji等[8]在领域语料和对比语料中分别计算TF等指标，并进一步进行融合；Bonin等[4]对N-Value、C-Value等指标进行整合，设计出一种适用于多单词领域术语评价指标；熊李艳等[15]应用背景语料的统计信息对C-Value指标进行改进；曾文等[7]针对科技文献的特点设计了基于词语组合强度和出现位置的领域术语统计指标。近年来，研究者开始将机器学习技术应用于领域术语识别问题，提出多种基于机器学习的方法，优势在于候选术语的各种特征可以被自动融合到识别模型中，避免人工指定规则和设计统计指标的困难，代表性工作有Foo等[16]利用语言模型提取领域术语特征；Da Silva Conrado等[17]设计了一套较为全面的用于训练术语分类模型的特征集；Loukachevitch[18]从搜索引擎查询结果中提取用于识别领域术语的特征。实验表明基于机器学习的领域术语识别方法能够取得较好的识别效果，但这类方法需要一定的人工标注，这制约了基于机器学习的领域术语识别方法的广泛应用。

本文提出的方法属于基于机器学习的方法。然而，本文并没有像传统方法一样应用监督学习算法，而是将领域术语识别任务抽象为一个多部图上的半监督学习问题，这种做法的优点是通过充分利用查询日志内在的流形结构以减少机器学习算法所需的人工标注量。此外，本文面向搜索引擎查询日志这种全领域语料，而已有基于机器学习的领域术语识别方法主要针对领域语料。

2.2 查询日志分析

搜索引擎查询日志是信息检索领域中的一种重要语料，除了具有海量、动态的特点之外，由于其中记录了海量用户与搜索引擎完整的交互过程，因而蕴含了丰富的“群体智慧”。近年来，在信息检索及相关领域开展了大量基于搜索引擎查询日志的研究工作，包括查询扩展[19]、、用户行为建模[20]、命名实体识别[21-24]等，其中与领域术语识别较为相关的是命名实体识别。命名实体是指诸如人名、机构名、地名等标识实体的名词，经统计发现，在搜索引擎查询日志中约70%的查询词包含命名实体[22]，因此查询日志成为命名实体识别的一个重要资源。目前代表性的命名实体识别方法有：翟海军等[21]使用弱监督主题模型识别命名实体的类别；Jain等[23]设计了一系列模式和统计指标从海量查询日志中高效识别领域无关的命名实体；Dalvi等[24]设计了面向命名实体识别任务的语言模型。

虽然命名实体与领域术语具有一定相似性，然而两者具有本质的不同：一方面，命名实体本质上是各种名称，故通常是名词，而领域术语并没有这个限制，比如“插秧”是农业领域的专业术语，但却不是一个命名实体；另一方面，命名实体与特定领域没有必然的关联性，而领域术语一定是与特定领域密切相关的。此外，已有工作很少基于查询日志内在的流形结构开展命名实体识别研究。

3 领域术语识别任务定义

3.1 查询日志形式化描述

查询日志记录了用户与搜索引擎交互过程中产生的各类信息，表1给出了查询日志中每条记录的基本字段格式，本文主要关注User、Query、URL三个字段。

表1 日志文件格式

字段	日志记录内容
TimeStamp	用户提交查询的时间
User	用户ID
Query	用户提交的查询词
URL	用户点击的URL地址
ShowRank	URL在搜索引擎中返回结果中的排名
ClickRank	用户点击URL的顺序

查询日志记录了搜索用户(User)、查询词(Query)、目标网页(URL)之间的关联关系。通过抓取URL 对应的网页并提取其中的候选术语，可以进一步得到 URL 与 Term 之间的关联关系。本文使用图 1 所示的四部图抽象描述上述各种关联关系。假设Q、U、P分别表示查询日志中所有查询词、用户、URL的集合，T表示 URL 对应网页中所有候选术语的集合，则查询日志四部图可表示为 G=<U∪Q∪P∪T, $E_{UQ}\cup E_{QP}\cup E_{PT}$>，其中$E_{UQ}$、$E_{QP}$、$E_{PT}$分别表示各类节点之间的边集。如果用户u提交了查询词q，那么u和q之间在E_{UQ}中存在连边；如果某用户通过查询词q点击了 URL p，那么q和p之间在E_{QP}中存在连边；如果URL 中对应的网页p中包含候选术语t，那么p和t之间在E_{PT}中存在连边。

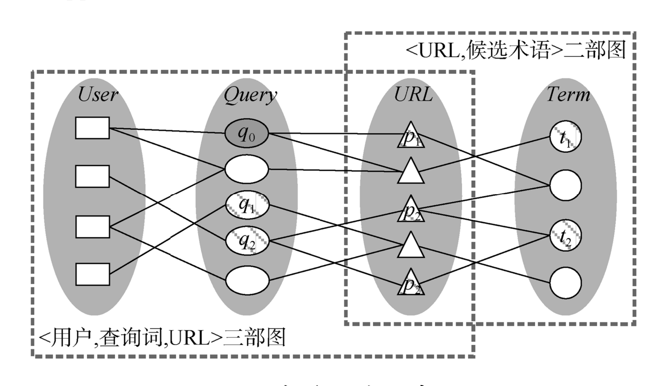


图 1　查询日志四部图

3.2　领域术语识别框架

基于查询日志四部图G，领域术语识别的目标是根据用户指定的初始领域查询词集合Q_0得到该领域内的术语集合T_D，本文将之分解为两个阶段实现：

(1) 领域查询词识别。给定目标领域，在 G 的<用户,查询词,URL>三部子图G_{UQP}=<U∪Q∪P, $E_{UQ}\cup E_{QP}$>中，识别与Q_0(图 1 中的q_0)具有相同领域的查询词集合Q_D(图 1 中的q_1和q_2)；

(2) 领域术语识别。领域查询词并不能被严格地作为领域术语，主要原因是用户输入的查询词在语法或词法格式上并不规范，口语化、冗长及歧义等现象普遍存在，因而基于Q_D，提取与之相关的网页集合P_D(图1中的p_1、p_2、p_3)，进一步在G的<URL,候选术语>二部子图G_{PT}=<P_D∪T, E_{PT}>中识别与目标领域相关的术语集合T_D(图 1 中的t_1和t_2)。

4　领域术语识别算法

由于查询日志涵盖众多领域，导致任一单个领域相关的信息均非常稀疏。领域稀疏性导致在查询日志四部图中识别领域查询词和领域术语变得非常困难。本文通过利用查询日志内在的结构化特征来克服稀疏性问题。由于同一用户通常具有相对专一的查询兴趣，且同一 URL 对应的网页通常是相对集中的主题，因而如果多个查询词经常被同一用户提交或者触发了相同URL 的点击，那么它们更倾向于属于同一领域。此外，领域术语具有聚集出现的特点，如果多个术语经常在一个网页中出现，那么它们也倾向于属于相同领域。基于上述特点，本文首先度量查询日志中查询词、用户、URL、候选术语之间的关联关系，得到G中各类边的权重，再应用流形排序算法识别与目标领域相关的查询词或术语。

4.1　领域查询识别

(1) <用户, 查询词, URL>三部子图边权重计算

三部子图G_{UQP}中边权重反映用户与查询词、查询词与 URL 之间的关联度。直观而言，如果用户u频繁提交查询词q，而且u提交的查询词总数较少，那么u与q之间具有较强关联度；如果 URL p频繁通过查询词q被点击，而且与p关联的查询词总数较少，那么p与q之间具有较强关联性。基于上述思想，对于∀< u,q>∈E_{UQ}及∀< q,p>∈E_{UQ}，本文设计如下权重公式：

$$w_{<u,q>} = \frac{\text{count}(u,q)}{\sum_{u'\in U}\text{count}(u',q)} \cdot \log\frac{|Q|}{\sum_{q'\in Q} I(u,q')} \tag{1}$$

$$w_{<q,p>} = \frac{\text{count}(p,q)}{\sum_{p'\in P}\text{count}(p',q)} \cdot \log\frac{|Q|}{\sum_{q'\in Q} I(p,q')} \tag{2}$$

其中，count(u,q)和count(p,q)分别表示在查询日志中用户u提交查询词q以及URL p通过查询词q被点击的次数，I(u,q')和I(p,q')分别表示用户u是否提交过查询词q'以及URL p是否通过查询词q被点击。公式(1)和公式(2)右侧前半部分是特定查询词上的用户/URL频率，反映了用户/URL与该查询词的相关度；公式(1)和公式(2)右侧后半部分是用户/URL频率的倒数，反映了用户/URL的专注度，与特定查询词无关。当用户/URL与某查询词之间相关度较高，且用户/URL的专注度较高时，则它与该查询词之间有较高关联度。

(2) 流形排序算法应用

流形排序算法[25]目标是利用图数据内在流形结构对图中每个节点进行排序。该算法在样本集上构建一个加权近邻图，人工赋予图中部分节点初始兴趣度；之后每个节点的分值在加权近邻图中线性迭代传播，直到达到一个稳态；排序在前的节点对应着具有较高兴趣度的样本。大量实践证明，流形排序收敛性较好，排序结果通常能有效反映图中各节点的兴趣度。

在领域查询词识别任务中，由领域专家指定一个初始领域查询词集合，在G_{UQP}上应用流形排序算法得到查询日志中所有查询词的排序，其中排名靠前的查询词与初始指定的领域查询词具有较强相关性，可被认为是领域查询词。由于G_{UQP}是异质图，在应用流形排序算法时需要先将其转换为由查询词节点构成的同质图。为此，需要基于G_{UQP}中各类边的权重计算得到在用户和 URL 维度上查询词之间的相似度，根据余弦相似度[26]设计了如下查询词相似度计算公式：

$$sim_{User}(q_i,q_j)=\frac{\sum_{u\in U} w_{<u,q_i>}\cdot w_{<u,q_j>}}{\sqrt{\sum_{u\in U} w^2_{<u,q_i>}}\cdot\sqrt{\sum_{p\in P} w^2_{<u,q_j>}}} \quad (3)$$

$$sim_{URL}(q_i,q_j)=\frac{\sum_{p\in P} w_{<q_i,p>}\cdot w_{<q_j,p>}}{\sqrt{\sum_{p\in P} w^2_{<q_i,p>}}\cdot\sqrt{\sum_{p\in P} w^2_{<q_j,p>}}} \quad (4)$$

之后，对用户和URL维度上查询词相似度进行线性加权融合，最终使用的查询词相似度公式如下：

$$sim(q_i,q_j)=\alpha\cdot sim_{User}(q_i,q_j)+(1-\alpha)\cdot sim_{URL}(q_i,q_j) \quad (5)$$

其中，α是参数，控制了用户维度和 URL 维度对查询词相似度贡献的比例。由于网页主题相对用户搜索兴趣更专一，因而 URL 维度的贡献通常更大。实验中令$\alpha=0.2$。

根据公式(5)可构建一个由查询词构成的查询词图，每个节点是一个查询词，节点间连边的权重由公式(5)计算。由于$sim(q_i,q_i)=1$ $(i=1,\cdots,|Q|)$，这会导致流形算法在运行过程中出现自加强现象，故令$\forall i=1,\cdots,|Q|$, $sim(q_i,q_i)=0$。

假设查询词图邻接矩阵$W=(sim(q_i,q_j))_{|Q|\times|Q|}$，在其上应用流形排序算法识别领域查询词的步骤如下：

①查询词图邻接矩阵预处理。为保证流形排序算法的收敛性，对W在行和列上进行归一化[18]，得到矩阵：

$$S=D^{-1/2}WD^{-1/2} \quad (6)$$

其中，D是对角阵，满足$D_{ii}=\sum_{j=1}^{|Q|}sim(q_i,q_j)$。

②领域度初始化。人工指定一个初始领域查询词集合，并定义向量$\mathbf{y}=(y_0,y_1,\cdots,y_{|Q|})^T$表示查询词节点的先验领域度，若查询词$q_i$属于初始领域查询词集合，$y_i=1$；否则$y_i=0$。

③领域度传播。定义向量$\mathbf{f}=(f_0,f_1,\cdots,f_{|Q|})^T$表示查询词节点的后验领域度，基于流行排序算法[25]进行迭代计算直到收敛，公式如下：

$$\mathbf{f}^{(t+1)}=\alpha\cdot S\cdot\mathbf{f}^{(t)}+(1-\alpha)\cdot\mathbf{y} \quad (7)$$

其中，$\alpha\in[0,1)$是平滑参数，用于控制先验领域度和相邻节点对最终领域度的贡献比例。

④领域查询词输出。指定领域查询词数量n，对每个查询词q_i $(i=1,\cdots,|Q|)$，根据f_i进行排序，取前n个查询词作为领域查询词Q_D。

4.2 领域术语识别

(1) 候选术语生成

识别得到领域查询词集Q_D后，在查询日志中用户通过Q_D点击的 URL 对应的网页即可被认为是领域相关的。本文将该领域网页集形式化表示为：

$$P_D=\{p\mid(\forall q\in Q_D)\wedge(p\in Click(q))\} \quad (8)$$

其中，Click(q)表示在查询日志中用户提交查询词q后点击的 URL 集合。抓取P_D中每个 URL 对应的网页，过滤 HTML 标签等无效信息后可得到该领域的 Web 语料集。由于抓取的网页是用户提交查询词后进行相关性判断后点击的，因而与领域查询词具有很高相关性，其中所有短语即为目标领域的候选术语集合。

然而，很多领域术语往往无法被传统中文分词工具正确识别，在很多情况下领域术语会被切分成分散的单字。比如农业领域术语“双条杉天牛”常被分词工具切分为“双/条/杉/天牛”。因此，对 Web 语料集进行切分后，需对切词结果进行合并以得到候选领域术语。具体使用基于滑动窗口的方法提取候选术语，即使用长度分别为 2、3、4 的滑动窗口得到分词词串中所有可能的 n-gram 作为可能的候选术语。由于 n-gram 数量庞大，本文使用扩展的多元互信息指标[27]度量候选 n-gram 的紧密度。

假设 n-gram $C=c_1c_2\cdots c_n$, $p(C)$ 和 $p(c_i)$ 分别为 n-gram C和词单元c_i在语料中出现的频率，本文为 n-gram C设计如下互信息公式。

$$\mathrm{eMI}(C)=\log\frac{p(C)}{\sqrt[\gamma]{\prod_{i=1}^{n}p(c_i)}} \tag{9}$$

相比于传统的多元互信息指标[27]，公式(9)中增加了开方参数γ。当γ取较大值时，可以对低频 n-gram 进行惩罚，降低低频 n-gram 的权重，从而排除一定的噪音低频 n-gram。实验中令$\gamma=2$。

由于候选术语包含不同长度的 n-gram，而不同长度的 n-gram 的互信息不具有可比性(互信息取值具有随n变大而增长的趋势)。为克服该问题，进一步在相同长度的 n-gram 之间进行归一化，得到作为最终度量 n-gram 紧密度的指标，公式如下：

$$\overline{\mathrm{eMI}}(C)=\mathrm{eMI}(C)\Big/\frac{1}{S_{|C|}}\sum_{C'\in S_{|C|}}\mathrm{eMI}(C') \tag{10}$$

其中，$S_{|C|}$表示与C具有相同长度的 n-gram 集合，因此分母表示所有与C具有相同长度的 n-gram 的互信息的平均值。使用公式(10)计算得到所有 n-gram 的紧密度后，紧密度超过某指定阈值的 n-gram 被选作候选术语。

(2) <URL, 候选术语>二部子图边权重计算

为了从候选术语中识别出真正的领域术语，先计算查询日志四部图的<URL, 候选术语>二部子图G_{PT}中候选术语与 URL 连边的权重，在其上应用流形排序算法得到所有候选术语关于领域度的排序。

G_{PT}中候选术语与 URL 连边的权重反映了两者的关联度，类似于公式(1)和公式(2)的设计思想，本文设计了如下连边权重计算公式：

$$w_{<p,t>}=\frac{\mathrm{count}(p,t)}{\mathrm{len}(p)}\cdot\log\frac{|P_D|}{\sum_{p'\in P_D}I(p',t)} \tag{11}$$

其中，count(p,t)表示候选术语 t 在网页p中出现的次数，len(p)表示网页p中候选术语数量，I(p,t)表示候选术语 t 是否在网页p中出现。

在得到G_{PT}的边权重之后，先将其转换为候选术语同质图，并在其上应用流形排序算法得到所有候选术语关于领域度的排序，序列中前m个候选术语即可被作为目标领域的术语集合。需要指出的是，由于在第一阶段已经得到领域查询词集合，对所有领域查询词进行切分，其中的高频词当作第二阶段所需的初始领域术语集合，而不再进行人工指定。

5　实验结果及分析

5.1　实验设置

实验使用某商用搜索引擎公开的真实查询日志①。为提高实验效率，笔者过滤掉无效的 URL 以及与之关联的查询词，最终用于实验的查询日志的统计信息如表 2 所示：

表 2　实验数据统计信息

统计项目	数量
查询词	3 400 480
URL	6 179 488
<查询词, URL>点击记录	6 905 367

实验选择“农作物病虫害防治”领域作为目标领域。为了能够对该领域的外延和内涵进行准确界定，笔者邀请一名某农业大学植物保护专业的博士参与实验，负责标注初始领域查询词以及人工评判算法输出的每条领域术语的正确性。

在领域术语识别任务中，目标领域内的领域术语通常难以全部获取，从而很难评价实验结果的召回率，因此本文采用基于精确率的评价指标。由于流形排序算法最终得到的是所有候选术语关于领域度的排序，具体采用评价排序准确性的指标 Precision@n(简称P@n)评价算法输出结果的准确度：

$$P@n=\frac{\text{输出序列的前n个术语为正确领域术语的数量}}{n} \tag{12}$$

其中，n 为在输出序列中截取的序列长度，本实验中令 n=100, 200, 300, 400, 500。

在领域术语识别研究中，使用查询日志作为语料的工作较少，文献[11]是其中的少数工作之一，故以文献[11]的方法作为本实验的基准方法。文献[11]需要事先人工标注一定数量的领域 URL 来获取领域 Web 语料，而本文方法仅需要人工指定少量初始领域查询词。为进行公平全面的比较，在实验中使用本文方法第一阶段中得到的领域查询词对应的 URL 替代文献[11]中所需人工标注的领域 URL。

①http://www.sogou.com/labs/dl/q.html.

5.2 实验结果

(1) 参数敏感性分析

对流形排序算法中平滑参数α的敏感性进行实验分析。首先，令α在[0.5, 1.0]范围内以 0.1 为步长取不同的值，并记录不同α取值下领域术语识别结果的 P@n 变化情况，结果如图 2 所示。可以看到，参数 α 对领域术语识别效果具有一定影响。随着α的增长，领域术语识别效果不断提升，当α=0.9时达到极值。由于α控制了先验领域度和相邻节点对最终领域度的贡献比例，这说明相比于人工指定的先验领域度，相邻节点间的领域度传播对于领域术语识别具有更重要的作用。然而，当α=1.0时，领域术语识别效果大大降低，这说明人工指定的先验领域度也是必不可少的。

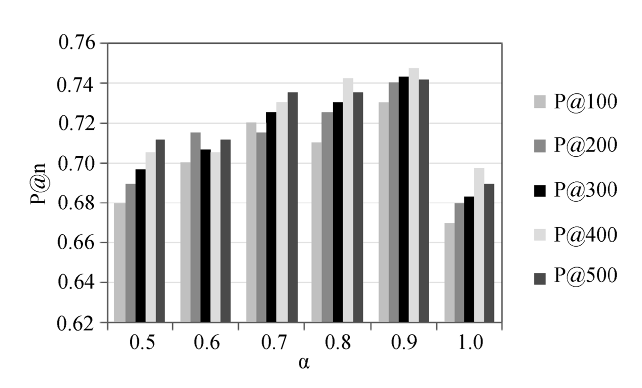


图 2 α不同取值下的领域术语识别效果

之后，令α=0.9，对本文方法两个阶段中应用流形排序算法时的收敛性能进行分析。实验中，分别计算相邻两个迭代过程中各个候选术语得分差值的平方和(Sum of Squared Difference, SSD)，以此评价迭代过程的收敛情况，结果如图 3 所示。可以看出，本文方法两个阶段均只需少量的迭代次数即可达到收敛，在领域查询词识别阶段迭代 150 次即可达到收敛，在领域术语识别阶段迭代 100 次即可达到收敛。因而，本文方法具有较好的领域术语识别效率。

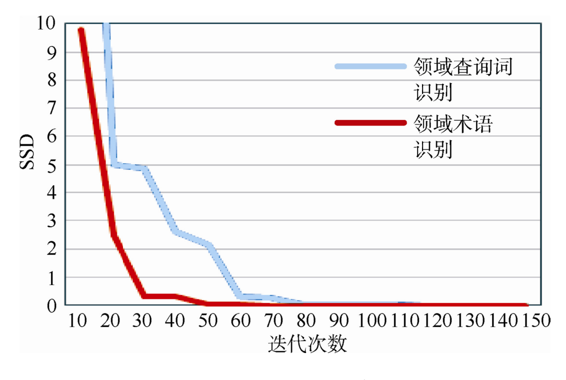


图 3 收敛性能

(2) 对比实验

基于上述参数取值，将本文方法与基准方法进行比较，分别标注两者识别得到的前 500 个领域术语的正确性，结果如图 4 所示。可以看出，本文方法的识别结果准确率平均达到 74%，具有一定的实用价值；此外，本文方法在各个 P@n 度量指标上都优于基准方法，特别在 n 取值较大时具有更大的优势，这说明相比于通过计算传统的领域度指标识别领域术语，通过构建查询日志四部图并应用流形排序算法能取得更优的领域术语识别效果。

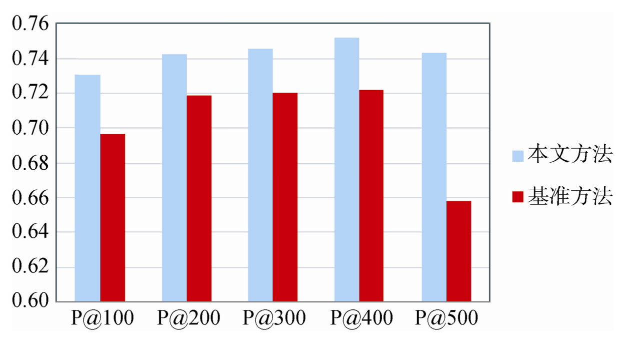


图 4 对比实验结果

5.3 识别结果展示及错误分析

表 3 展示了本文方法在农作物病虫害防治领域中识别得到的部分领域术语，其中第 1 列是人工指定的初始领域查询词，第 2 列和第 3 列是识别得到的前 30 个领域术语及其排序得分，其中标"*"的是识别错误。可以发现，本文方法在整体上能够得到较为准确的识别结果。

通过对识别错误进行分析，发现本文方法存在两方面的局限性：如果某些频率较高的词与初始查询词频繁共现，那么倾向于将这些组合搭配词识别为领域术语，例如"玉米螟的"、"蚜虫蚜虫"，该问题可以通过进一步改进候选术语紧密度公式解决；识别得到的领域术语大多与指定的种子术语相关性高，若种子术语对领域内容的覆盖面不足，则有可能导致识别结果的召回率不高，为解决该问题，需要根据目标领域的知识类目进行有代表性的选取，比如对于农作物病虫害防治领域，可以参考农业图谱等领域知识资源，并邀请相关专业(植物保护等)的专家指定高质量的初始领域查询词。

表 3 领域词识别结果

序号	初始领域查询词	序号	领域术语	领域度	序号	领域术语	领域度
1	蚜虫	1	白粉病	1.0000	16	卷叶螟	0.3881
2	蓟马	2	褐斑病	1.0000	17	敌敌畏	0.3754
3	蛴螬	3	玉米螟	1.0000	18	雌成虫	0.2859
4	白粉病	4	菌丝体	1.0000	19	果实膨大期	0.2845
5	疫病	5	蚜虫	1.0000	20	草莓白粉病	0.2818
6	乳油	6*	玉米螟的	0.8324	21	菊花褐斑病	0.2776
7	介壳虫	7	波尔多液	0.8061	22	苦瓜褐斑病	0.2752
8	菌丝体	8	多菌灵	0.6676	23	芍药褐斑病	0.2739
9	褐斑病	9	炭疽病	0.6592	24	苦瓜白粉病	0.1998
10	敌敌畏	10	功夫乳油	0.5468	25*	蚜虫蚜虫	0.1838
11	根腐病	11	青枯病	0.5101	26	橡胶白粉病	0.1684
12	炭疽病	12	白色菌丝体	0.4811	27	番木瓜白粉病	0.1633
13	吹绵蚧	13	絮状菌丝体	0.4433	28	油菜蚜虫	0.1625
14	白绢病	14	菌丝体迅速	0.4015	29	豌豆蚜虫	0.1579
15	玉米螟	15	茎基部	0.3990	30	冬瓜白粉病	0.1451

6 结 语

搜索引擎查询日志是一种重要的语料资源，具有海量、动态等特点，其中蕴含丰富的领域术语。本文提出利用查询日志自动识别领域术语的方法。在该方法中，查询日志被抽象成一个四部图结构，通过在其上应用流形排序算法分别得到候选术语关于领域度的排序，在序列中排在前列的被认为是目标领域相关的领域术语。本文方法的优点是：能够自动从查询日志这类非领域语料中识别出特定领域的领域术语，避免了传统方法需要事前准备大规模领域语料这一现实难题；通过充分利用查询日志内在的结构化特征，只需标注少量的初始领域查询词，即可识别得到丰富准确的领域术语，避免了传统方法所需的大量人工标注工作。在真实的查询日志数据集上进行实验，结果表明本文方法具有较高的收敛速度和识别准确率。进一步研究工作包括引入半监督学习和主动学习机制，进一步降低领域术语识别方法对初始领域查询词的依赖。

参考文献:

[1] 刘春燕，安小米，侯人华. 术语标准研制方法及在信息与文献领域中的应用[J]. 图书情报工作，2014，58(9): 91-95. (Liu Chunyan, An Xiaomi, Hou Renhua. Vocabulary Standard Development Methodology and Its Application in the Information and Documentation Fields [J]. Library and Information Service, 2014, 58(9): 91-95.)

[2] Caracciolo C, Stellato A, Morshed A, et al. The AGROVOC Linked Dataset [J]. Semantic Web, 2013, 4(3): 341-348.

[3] Bodenreider O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology [J]. Nucleic Acids Research, 2004, 32(S1): D267-D270.

[4] Bonin F, Dell'Orletta F, Venturi G, et al. A Contrastive Approach to Multi-word Term Extraction from Domain Corpora[C]. In: Proceedings of the 7th International Conference on Language Resources and Evaluation. 2010: 3222-3229.

[5] 化柏林. 针对中文学术文献的情报方法术语抽取[J]. 现代图书情报技术，2013(6):68-75. (Hua Bolin. Extracting Information Method Term from Chinese Academic Literature [J]. New Technology of Library and Information Service, 2013(6):68-75.)

[6] 何远标，乐小虬，张帆. 学术论文大纲中关键术语抽取方法研究[J]. 现代图书情报技术，2014(3): 73-79. (He Yuanbiao, Le Xiaoqiu, Zhang Fan. Research on Keyphrase Extraction from Scholarly Article Outline [J]. New Technology of Library and Information Service, 2014(3): 73-79.)

[7] 曾文，徐硕，张运良，等. 科技文献术语的自动抽取技术研究与分析[J]. 现代图书情报技术，2014(1):51-55. (Zeng Wen, Xu Shuo, Zhang Yunliang, et al. The Research and Analysis on Automatic Extraction of Science and Technology Literature Terms [J]. New Technology of Library and Information Service, 2014(1): 51-55.)

[8] Dorji T C, Atlam E-S, Tata S, et al. Extraction, Selection and Ranking of Field Association (FA) Terms from Domain-specific Corpora for Building a Comprehensive FA Terms Dictionary[J]. Knowledge and Information Systems, 2011, 27(1): 141-161.

[9] 屈鹏，王惠临. 面向信息分析的专利术语抽取研究[J]. 图书情报工作，2013, 57(1):130-135. (Qu Peng, Wang Huilin. Patent Term Extraction for Information Analysis [J]. Library and Information Service, 2013, 57(1): 130-135.)

[10] 谷俊, 王昊. 基于领域中文文本的术语抽取方法研究[J]. 现代图书情报技术, 2011(4): 29-34. (Gu Jun, Wang Hao. Study on Term Extraction on the Basis of Chinese Domain Texts [J]. New Technology of Library and Information Service, 2011(4):29-34.)

[11] 闫兴龙, 刘奕群, 方奇, 等. 基于网络资源与用户行为信息的领域术语提取[J]. 软件学报, 2013, 24(9):2089-2100. (Yan Xinglong, Liu Yiqun, Fang Qi, et al. Domain-Specific Terms Extraction Based on Web Resource and User Behavior [J]. Journal of Software, 2013, 24(9): 2089-2100.)

[12] Jiang D, Pei J, Li H. Mining Search and Browse Logs for Web Search: A Survey [J]. ACM Transactions on Intelligent Systems and Technology, 2013, 4(4): Article No. 57.

[13] 季培培, 鄢小燕, 岑咏华. 面向领域中文文本信息处理的术语识别与抽取研究综述[J]. 图书情报工作, 2010, 54(16): 124-129. (Ji Peipei, Yan Xiaoyan, Cen Yonghua. A Survey of Term Recognition and Extraction for Domain-specific Chinese Text Information Processing [J]. Library and Information Service, 2010, 54(16): 124-129.)

[14] 宋培彦, 路青, 刘宁静. 一种从术语定义句中自动抽取知识单元的方法[J]. 情报杂志, 2014, 33(4): 139-143. (Song Peiyan, Lu Qing, Liu Ningjing. A New Method for Knowledge Unit Automatic Extraction Using Definitions of Terms [J]. Journal of Intelligence, 2014, 33(4): 139-143.)

[15] 熊李艳, 谭龙, 钟茂生. 基于有效词频的改进 C-value 自动术语抽取方法[J]. 现代图书情报技术, 2013(9): 54-59. (Xiong Liyan, Tan Long, Zhong Maosheng. An Automatic Term Extraction System of Improved C-value Based on Effective Word Frequency [J]. New Technology of Library and Information Service, 2013(9): 54-59.)

[16] Foo J, Merkel M. Using Machine Learning to Perform Automatic Term Recognition [C]. In: Proceedings of the LREC 2010 Workshop on Methods for Automatic Acquisition of Language Resources and Their Evaluation Methods, Malta.2010: 49-54.

[17] Da Silva Conrado M, Pardo T, Rezende S O. A Machine Learning Approach to Automatic Term Extraction Using a Rich Feature Set [C]. In: Proceedings of NAACL HLT 2013 Student Research Workshop. 2013: 16-23.

[18] Loukachevitch N V. Automatic Term Recognition Needs Multiple Evidence [C]. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. 2012: 2401-2407.

[19] Jiang D, Leung K W T, Yang L, et al. Query Suggestion with Diversification and Personalization [J]. Knowledge-Based Systems, 2015, 89: 553-568.

[20] Rose D E, Levinson D. Understanding User Goals in Web Search [C]. In: Proceedings of the 13th International Conference on World Wide Web. ACM, 2004:13-19.

[21] 翟海军, 郭嘉丰, 王小磊, 等. 基于用户查询日志的命名实体挖掘[J]. 中文信息学报, 2010, 24(1): 71-76, 116. (Zhai Haijun, Guo Jiafeng, Wang Xiaolei, et al. Mining Named Entities from Query Logs [J]. Journal of Chinese Information Processing, 2010, 24(1): 71-76, 116.)

[22] Xu G, Yang S H, Li H. Named Entity Mining from Click-through Data Using Weakly Supervised Latent Dirichlet Allocation [C]. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009:1365-1374.

[23] Jain A, Pennacchiotti M. Domain-independent Entity Extraction from Web Search Query Logs[C]. In: Proceedings of the 20th International Conference Companion on World Wide Web. ACM, 2011:63-64.

[24] Dalvi B, Xiong C, Callan J. A Language Modeling Approach to Entity Recognition and Disambiguation for Search Queries [C]. In: Proceedings of the 1st International Workshop on Entity Recognition & Disambiguation. ACM, 2014: 45-54.

[25] Zhou D, Weston J, Gretton A, et al. Ranking on Data Manifolds [J]. Advances in Neural Information Processing Systems, 2004, 16: 169-176.

[26] Singhal A. Modern Information Retrieval: A Brief Overview [J]. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2001, 24(4): 35-43.

[27] Van de Cruys T. Two Multivariate Generalizations of Pointwise Mutual Information [C]. In: Proceedings of the Workshop on Distributional Semantics and Compositionality. Association for Computational Linguistics, 2011: 16-20.

作者贡献声明:

刘彤: 实现并改进研究方案, 撰写论文;
倪维健: 提出研究思路, 设计研究方案, 修改论文;
柳梅: 整理实验数据, 协同完成实验。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据[1-3]见期刊网络版 http://www.infotech.ac.cn; 支撑数据[4-8]由作者自存储, 可通过电子邮件向作者索取, E-mail:

niweijian@gmail.com。

[1] 刘彤，倪维健，柳梅. top-query.txt. 识别得到的领域查询词排序.

[2] 刘彤，倪维健，柳梅. top-term.txt. 识别得到的领域术语排序.

[3] 刘彤，倪维健，柳梅. convergence.txt. 两阶段流形排序算法收敛效果.

[4] 刘彤，倪维健，柳梅. querylog-pre.txt. 处理后的查询日志数据.

[5] 刘彤，倪维健，柳梅. query-user.txt. 查询词/用户关联矩阵.

[6] 刘彤，倪维健，柳梅. query-url.txt. 查询词/URL 关联矩阵.

[7] 刘彤，倪维健，柳梅. webcorpus.txt. 领域网页数据集.

[8] 刘彤，倪维健，柳梅. url-term.txt. URL/候选术语关联矩阵.

收稿日期: 2015-08-13
收修改稿日期: 2015-12-10

Identifying Terminology from Search Engine Query Logs

Liu Tong　Ni Weijian　Liu Mei
(College of Information Science and Engineering, Shandong University of Science and Technolgoy, Qingdao 266590, China)

Abstract: [**Objective**] This study proposes a new approach to identify terminologies from search engine query logs for the purpose of improving traditional technology. [**Methods**] First, used the four-partite graph to re-present those query logs. Then, ranked the candidate terminologies with the help of manifold ranking algorithm. Those top ranked ones were domain-specified. [**Results**] We tested the proposed method with real search engine query logs and found the precision rates were about 20% higher than the standard approach. [**Limitations**] The coverage of those identified terminologies relies on the initial domain-specified queries manually chosen by the experts. [**Conclusions**] The proposed approach could build high quality domain thesaurus without pre-defined large domain corpus and annotations. Thus, the new method was more practical for real world issues.

Keywords: Domain terminology　Search engine　Query logs　Manifold ranking